

Optimal Data and Training Symbol Ratio for Communication over Uncertain Channels

Ather Gattami

Ericsson Research

Stockholm, Sweden

Email: ather.gattami@ericsson.com

Abstract—We consider the problem of determining the power ratio between the training symbols and data symbols in order to maximize the channel capacity for transmission over uncertain channels with a channel estimate available at both the transmitter and receiver. The receiver makes an estimate of the channel by using a known sequence of training symbols. This channel estimate is then transmitted back to the transmitter. The capacity that the transceiver maximizes is the worst case capacity, in the sense that given a noise covariance, the transceiver maximizes the minimal capacity over all distributions of the measurement noise under a fixed covariance matrix known at both the transmitter and receiver. We give an exact expression of the channel capacity as a function of the channel covariance matrix, and the number of training symbols used during a coherence time interval. This expression determines the number of training symbols that need to be used by finding the optimal integer number of training symbols that maximize the channel capacity. As a bi-product, we show that linear filters are optimal at both the transmitter and receiver.

NOTATION

$\det(A)$	$\det(A) = \prod_i \lambda_i$, where $\{\lambda_i\}$ are the eigenvalues of the square matrix A .
\otimes	$A \otimes B$ denotes the kronecker product between the matrices A and B .
$\mathbf{E}\{\cdot\}$	$\mathbf{E}\{x\}$ denotes the expected value of the stochastic variable x .
$\mathbf{E}\{\cdot \cdot\}$	$\mathbf{E}\{x y\}$ denotes the expected value of the stochastic variable x given y .
cov	$\text{cov}\{x, y\} = \mathbf{E}\{xy^*\}$.
$h(x)$	Denotes the entropy of x .
$h(x y)$	Denotes the entropy of x given y .
$I(x; y)$	Denotes the mutual information between x and y .
$\mathcal{N}(m, V)$	Denotes the set of Gaussian variables with mean m and covariance V .

I. INTRODUCTION

A. Background

This work considers the problem of determining the power ratio between the training symbols and data symbols in order to maximize the channel capacity for transmission over uncertain channels with channel state information at the transmitter and receiver. While the problem of MIMO communication over a channel known at the transmitter and receiver is well understood, the problem of uncertain channels still needs a deeper understanding of how much power we should spend

on estimating the channel in order to transmit as much data as possible. This problem poses a trade-off between the power ratio allocated for training and data transmission. On one hand, if we spend more power on training and less on data transmission, the data throughput will be small of course. On the other hand, if we spend most of the power on data transmission and much less on training, the channel estimate will be bad and therefore, the channel estimation error noise will be large, causing a rather low data rate. Given *per symbol* power constraints, one might need a number of training symbols in order to achieve a certain quality of channel state information, which leaves a smaller number of symbols for data transmission. This is an important constraint that we take into account in this work.

B. Previous Work

There has been a lot of work on MIMO communication in the context of uncertain channel state information. The seminal paper of Telatar [7] studied the problem of communication over uncertain Gaussian channels under the assumption that the channel realization is available at the receiver but not the transmitter. However, in practice, the transmitter and receiver don't have full knowledge of the channel. The work was extended in [6] for the case of slowly varying channels. The channel estimation extension of [7] was presented in [3], where the problem of power ratio between the training and data symbols was studied under *average* and *per symbol* power constraints over the channel coherence time (the time where the channel is roughly constant). The crucial assumption of average power constraint allows for spending only one symbol on training, since the power is only limited by the total power resource available during the coherence time. For the case of *per symbol* power constraints, the power ratio is harder to compute. The case where a channel estimate is available at the transmitter and receiver was studied in [1] under the specific signaling strategy of zero-forcing linear beamforming and *given Gaussian measurement noise*. Calculation of the channel capacity with respect to a channel estimate available at the receiver only was given in [2].

C. Contribution

We consider communication over uncertain channels with channel state feedback at the transmitter. The receiver makes

an estimate of the channel by using a known sequence of training symbols. This channel estimate is then *transmitted back* to the transmitter. The capacity that the transceiver maximizes is the worst case capacity, in the sense that given a noise covariance, the transceiver maximizes the minimal capacity over all distributions of the measurement noise under a fixed covariance matrix known at both the transmitter and receiver. The channel estimation error implies that the covariance of the transmitted symbols over time affects both the covariance of the transmitted information symbols and the total noise covariance, which makes the signal structure more complicated, where it's not clear if the symbols should be uncorrelated over time. For the single input multiple output (SIMO) channel, we give an exact expression of the channel capacity as a function of the channel covariance matrix, the noise covariance matrix, and the number of training symbols used during a coherence time interval. This expression determines the number of training symbols that need to be used by finding the optimal integer number of training symbols that maximize the channel capacity. Numerical examples illustrate the trade-off between the number of training and data symbols. The results indicate that when the transmission power (or equivalently the signal to noise ratio) is high, a smaller number of training symbols is required to maximize the capacity compared to the low transmission power case. We confirm these observations theoretically considering the asymptotic behavior of the power as it grows large or decreases to very small values.

II. PRELIMINARIES

Definition 1 (Kronecker Product): For two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$, the Kronecker product is defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

Proposition 1 (Multiplication Property of the Kronecker Product): For any set of matrices A, B, C, D , we have that

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

Proof: Consult [4]. ■

Proposition 2 (Determinant Property): For any two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$, we have that

$$\det(AB + I_m) = \det(BA + I_n)$$

Proof: Consult [5]. ■

Proposition 3 (AM-GM Inequality): Let a_1, a_2, \dots, a_n be n nonnegative real numbers. Then

$$\frac{1}{n} \sum_{i=1}^n a_i \geq \sqrt[n]{\prod_{i=1}^n a_i}$$

with equality if and only if $a_1 = a_2 = \dots = a_n$.

Proof: The proof may be found in most standard textbooks on Calculus. ■

III. PROBLEM FORMULATION

Let H be a random channel such that $\text{vec}(H) \sim \mathcal{CN}(0, C)$. The realization of the channel is assumed to be constant over a coherence time interval corresponding to a block of T symbols. Let $x(t) \sim \mathcal{CN}(0, X)$ be the transmitted symbol over the Gaussian channel at time t . The received signal at time t is given by

$$y(t) = Hx(t) + w(t)$$

where $\{w(t)\}$ is random white noise process, independent of H and x , with zero mean and covariance given by $\mathbf{E}\{w(t)w^*(t)\} = I_m$ known at both the transmitter and receiver. Without loss of generality, we assume that W is invertible.

Let $(x(1), \dots, x(T))$ be the block symbols transmitted within the channel coherence time T . The *average* power constraint imposed on the *block* is given by

$$\frac{1}{T} \sum_{t=1}^T \mathbf{E}\{|x(t)|^2\} \leq P$$

The power constraint above could allow for some symbols $x(t)$ to have a larger power than P . However, in the real world, we have hard constraints on the peak average power per symbol, so we will impose power constraints per symbol given by

$$\mathbf{E}\{|x(t)|^2\} \leq P, \quad \text{for } t = 1, \dots, T$$

Suppose that we allocate T_τ training symbols for the channel estimation part and $T_d = T - T_\tau$ symbols for data transmission. If a performance criterion is measured over the total channel coherence time T , then clearly the optimal strategy is to transmit the training symbol sequence first followed by the data symbols. Also, when no energy constraints are present, the optimal power allocation is for each symbol to be transmitted with full power P . However, it's not clear what the optimal *ratio* between the training and data symbols. That is, what is the optimal choice of T_τ to maximize the channel capacity? To this end, we will derive the exact expression of the channel capacity SIMO channel.

A. SIMO Channel Estimation

Consider a random Gaussian channel H taking values in \mathbb{R}^m . Let x_τ be a deterministic training symbol known at the transmitter and receiver with $|x_\tau|^2 = P$. The transmitted training sequence is given by

$$x(t) = x_\tau, \quad \text{for } t = 1, \dots, T_\tau$$

At the receiver, we obtain the measurement symbols

$$y(t) = Hx_\tau + w(t), \quad \text{for } t = 1, \dots, T_\tau$$

Introduce the vectors

$$y_\tau = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(T_\tau) \end{pmatrix}, \quad w_\tau = \begin{pmatrix} w(1) \\ w(2) \\ \vdots \\ w(T_\tau) \end{pmatrix}$$

and let the covariance matrix of w_τ be

$$\mathbf{E}\{w_\tau w_\tau^*\} = I_m \otimes I_{T_\tau} = I_{mT_\tau}$$

It's well known that the optimal estimator \hat{H} of H given y_τ that minimizes the MSE is given by

$$\hat{H} = \mathbf{E}\{H | y_\tau\} \quad (1)$$

$$= \mathbf{cov}(H, y_\tau) \cdot \mathbf{cov}(y_\tau, y_\tau)^{-1} y_\tau \quad (2)$$

$$= \mathbf{E}\{H y_\tau^*\} \cdot \mathbf{E}\{y_\tau y_\tau^*\}^{-1} y_\tau \quad (3)$$

$$= x_\tau^* C \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} \frac{1}{T_\tau} \sum_{t=1}^{T_\tau} y(t) \quad (4)$$

$$= x_\tau^* C \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} (H x_\tau + \bar{w}(t)) \quad (5)$$

where

$$\bar{w}(t) = \frac{1}{T_\tau} \sum_{t=1}^{T_\tau} w(t)$$

The covariance of $\bar{w}(t)$ is easily obtained from the expression above, and it's given by

$$\mathbf{E}\{\bar{w}(t) \bar{w}^*(t)\} = \frac{1}{T_\tau} I_m$$

The channel estimation error is given by

$$\tilde{H} = H - \hat{H} \quad (6)$$

$$= H - x_\tau^* C \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} (H x_\tau + \bar{w}(t)) \quad (7)$$

$$= (I_m - PC \left(PC + \frac{1}{T_\tau} I_m \right)^{-1}) H - x_\tau^* C \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} \bar{w}(t) \quad (8)$$

$$= \frac{1}{T_\tau} \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} H - x_\tau^* C \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} \bar{w}(t) \quad (9)$$

Now we have that

$$\hat{C} = \mathbf{E}\{\hat{C}\} \quad (10)$$

$$= PC \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} C \quad (11)$$

It's well known that \hat{H} and \tilde{H} are independent since H and y_τ are jointly Gaussian. Thus,

$$\tilde{C} = \mathbf{E}\{\tilde{H} \tilde{H}^*\} \quad (12)$$

$$= \mathbf{E}\{(H - \hat{H})(H - \hat{H})^*\} \quad (13)$$

$$= C - \hat{C} \quad (14)$$

$$= C - PC \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} C \quad (15)$$

$$= (I_m - PC \left(PC + \frac{1}{T_\tau} I_m \right)^{-1}) C \quad (16)$$

$$= \frac{1}{T_\tau} \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} C \quad (17)$$

B. Channel Capacity

In this section, we will derive a formula for the channel capacity under the assumption that the transmitted and receiver have a common estimate of the channel. The estimate is the optimal estimate that minimizes the expected value of the variance of the estimation error.

Suppose that we transmit T_τ training symbols during the time interval $t = 1, \dots, T_\tau$ and $T_d = T - T_\tau$ data symbols during the time interval $t = T_\tau + 1, \dots, T$. The received noisy measurements of the data symbols are given by

$$y(t) = H x_d(t) + w(t) \quad (18)$$

$$= \hat{H} x_d(t) + \tilde{H} x_d(t) + w(t) \quad (19)$$

$$= \hat{H} x_d(t) + v(t), \quad (20)$$

for $t = T_\tau + 1, \dots, T$. Note that $v(t) = \tilde{H} x_d(t) + w(t)$ is uncorrelated with $x_d(t)$ and $\hat{H} x_d(t)$ jointly. Introduce the vectors

$$y_d = \begin{pmatrix} y(T_\tau + 1) \\ y(T_\tau + 2) \\ \vdots \\ y(T) \end{pmatrix}, \quad x_d = \begin{pmatrix} x_d(T_\tau + 1) \\ x_d(T_\tau + 2) \\ \vdots \\ x_d(T) \end{pmatrix}$$

$$w_d = \begin{pmatrix} w(T_\tau + 1) \\ w(T_\tau + 2) \\ \vdots \\ w(T) \end{pmatrix}, \quad v_d = \begin{pmatrix} v(T_\tau + 1) \\ v(T_\tau + 2) \\ \vdots \\ v(T) \end{pmatrix}$$

and let the respective covariance matrices of w_d and v_d be

$$W_d = \mathbf{E}\{w_d w_d^*\} = I_m \otimes I_{T_d} = I_{mT_d}$$

and

$$V_d = \mathbf{E}\{v_d v_d^*\} = V \otimes I_{T_d} = \begin{pmatrix} V & 0 & \cdots & 0 \\ 0 & V & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V \end{pmatrix}$$

The *worst case* capacity of the channel with respect to the measurement noise I_{mT_d} is given by

$$\mathcal{C}(T_\tau) = \sup_{x_d} \inf_{v_d} \mathbf{I}(x_d; y_d) \quad (21)$$

$$\mathbf{E}\{|x_d(t)|^2\} = P \quad \mathbf{E}\{v_d v_d^*\} = I_{mT_d}$$

Thus, the problem that we want to solve is as follows.

Problem 1: Find the optimal integer $T_\tau \in [1, T]$ such that

$$\mathcal{C}(T_\tau) = \sup_{\substack{x_d \\ \mathbf{E}\{|x_d(t)|^2\} = P}} \inf_{\substack{v_d \\ \mathbf{E}\{v_d v_d^*\} = I_{mT_d}}} \mathbf{I}(x_d; y_d)$$

is maximized.

IV. MAIN RESULTS

It's well known that for a deterministic channel \hat{H}_d and signal measurement

$$y_d = \hat{H}_d x_d + v_d$$

with measurement noise v_d uncorrelated with the transmitted signal x_d , the optimal transmitting strategy is for x_d to be Gaussian in order to minimize the worst case noise v_d which is also shown to be Gaussian. In other words, the Gaussian input and noise form a Nash equilibrium. The case when the noise v_d has an *arbitrary* covariance matrix V_d has been solved in [3]. In our case, the situation is different. The covariance matrix of v_d has a *structure* that also depends on the choice of the covariance matrix of the transmitted signal x_d . More precisely, recall that we have assumed that $\{w(t)\}$ is a temporally uncorrelated noise process with arbitrary distribution. Introduce

$$\hat{H}_d = \hat{H} \otimes I_{T_d} = \begin{pmatrix} \hat{H} & 0 & \cdots & 0 \\ 0 & \hat{H} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{H} \end{pmatrix}$$

and

$$\tilde{H}_d = \tilde{H} \otimes I_{T_d} = \begin{pmatrix} \tilde{H} & 0 & \cdots & 0 \\ 0 & \tilde{H} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{H} \end{pmatrix}$$

We may write $w_d = v_d - \tilde{H}_d x_d$, with v_d uncorrelated with \hat{H}_d and x_d . Furthermore,

$$\mathbf{E}\{v_d v_d^*\} - \mathbf{E}\{\tilde{H}_d x_d x_d^* \tilde{H}_d^*\} = \mathbf{E}\{w_d w_d^*\} = I_{mT_d}$$

Let $X_d = \mathbf{E}\{x_d x_d^*\}$. Then we have

$$\hat{H}_d x_d = \begin{pmatrix} \hat{H} x_d(T_\tau + 1) \\ \hat{H} x_d(T_\tau + 2) \\ \vdots \\ \hat{H} x_d(T) \end{pmatrix}$$

and the blocks of $\hat{H}_d x_d x_d^* \hat{H}_d^*$ at position (i, j) will be given by (since $x_d(T_\tau + i)$ is a scalar for all $i = 1, \dots, T - T_\tau$)

$$\begin{aligned} [\hat{H}_d x_d x_d^* \hat{H}_d^*]_{ij} &= \hat{H} x_d(T_\tau + i) x_d^*(T_\tau + j) \hat{H}^* \\ &= x_d(T_\tau + i) x_d^*(T_\tau + j) \hat{H} \hat{H}^*, \end{aligned}$$

for $i, j = 1, \dots, T - T_\tau$. Thus, $\hat{H}_d x_d x_d^* \hat{H}_d^* = (x_d x_d^*) \otimes (\hat{H} \hat{H}^*)$ and given \hat{H} , we get

$$\mathbf{E}\{\tilde{H}_d x_d x_d^* \tilde{H}_d^*\} = X_d \otimes \hat{C}.$$

Similarly, we get $\tilde{H}_d x_d x_d^* \tilde{H}_d^* = (x_d x_d^*) \otimes (\tilde{H} \tilde{H}^*)$ and

$$\mathbf{E}\{\tilde{H}_d x_d x_d^* \tilde{H}_d^*\} = X_d \otimes \tilde{C}.$$

This gives in turn

$$V_d = \mathbf{E}\{v_d v_d^*\} = I_{mT_d} + X_d \otimes \tilde{C}$$

In particular, the block diagonal elements of V_d are equal, with the block elements given by

$$V = I_m + \mathbf{E}\{\tilde{H} P \tilde{H}^*\} = I_m + P \tilde{C}$$

Now the cost given by (21) may be written in terms of v_d instead. That is,

$$\mathcal{C}(T_\tau) = \sup_{\mathbf{E}\{|x_d(t)|^2\} = P} \inf_{\substack{v_d \\ \mathbf{E}\{v_d v_d^*\} = I_{mT_d} + X_d \otimes \tilde{C}}} \mathbf{I}(x_d; y_d) \quad (22)$$

Theorem 1: Consider a communication channel given by $y(t) = Hx(t) + w(t)$ with H taking values in \mathbb{R}^m and $H \sim \mathcal{CN}(0, C)$, $t = 1, \dots, T - T_\tau$. Let \hat{H} be the channel estimate that is available at both the transmitter and receiver, based on T_τ training symbols. Under the power constraint $\mathbf{E}\{x^2(t)\} \leq P$, the worst case capacity $\mathcal{C}(T_\tau)$ is given by

$$\mathcal{C}(T_\tau) = (T - T_\tau)(\log_2 \det(PC + I_m) - \log_2 \det(P\tilde{C} + I_m))$$

with

$$\tilde{C} = \frac{1}{T_\tau} \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} C$$

Furthermore, the optimal receiver is given by the linear estimator $\mathbf{E}\{x_d | \hat{H} x_d + v_d\}$ with $v_d \sim \mathcal{CN}(0, P\tilde{C})$.

Proof: The proof is deferred to the appendix. ■

Theorem 1 can be checked for the extreme cases of $T_\tau = 0$ and $T_\tau = T$. For the case $T_\tau = 0$, clearly no channel estimation is obtained and the expression of \tilde{C} reveals that this error becomes infinite. Thus, the capacity becomes $-\infty$ and no information may be transmitted. The other extreme, $T_\tau = T$, means that all power is spent on channel estimation and no data may be transmitted. We see that the expression of the channel capacity formula gives zero capacity, agreeing with the physical model.

V. NUMERICAL EXAMPLES

Example 1: Consider a communication SIMO channel H with perfect feedback and a randomly generated covariance matrix

$$C = \begin{pmatrix} 0.7426 & -0.7222 \\ -0.7222 & 6.4075 \end{pmatrix}$$

The length of the block is assumed to be $T = 100$. Figure 1 shows that the optimum number of training symbols is $T_\tau = 4$ when the power constraint is given by $P = 100$, whereas for $P = 0.01$, Figure 2 shows that the capacity is maximized for $T_\tau = 27$, which is 27% of the available transmission power.

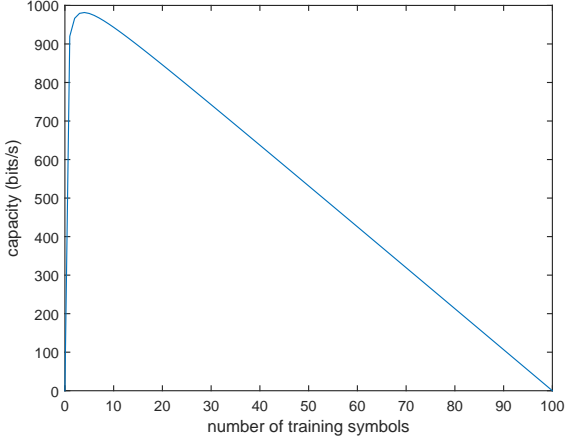


Fig. 1. A plot of the 2×1 channel capacity as a function of the number of training symbols with transmission power given by $P = 100$.

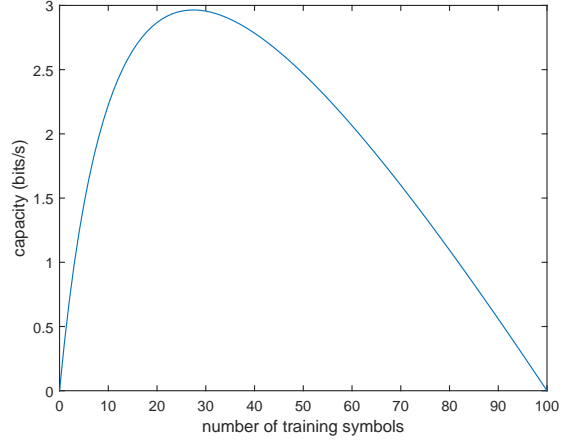


Fig. 2. A plot of the 2×1 channel capacity as a function of the number of training symbols with transmission power given by $P = 0.01$.

Example 2: Consider a communication SIMO channel H with perfect feedback and a randomly generated covariance matrix C given in the appendix. The length of the block is assumed to be $T = 100$. Figure 3 shows that the optimum number of training symbols is $T_\tau = 2$ when the power constraint is given by $P = 100$, whereas for $P = 0.01$, Figure 4 shows that the capacity is maximized for $T_\tau = 19$. The example clearly shows that the number of pilots needed is smaller when the number of receiving antennas increases. This is because of the increase of the received signal power which makes the system less sensitive to channel estimation errors.

VI. ASYMPTOTIC RESULTS

The previous numerical examples clearly show the dependence of the number of training symbols on the symbol transmit power. We may understand this relation by looking at the asymptotic results when the power P tends to 0 or infinity (low respectively high signal to noise ratio). Recall that the covariance of the channel estimation error is given by

$$\tilde{C} = \frac{1}{T_\tau} \left(PC + \frac{1}{T_\tau} I_m \right)^{-1} C$$

Clearly, if P is very small and T_τ is small, then $\frac{1}{T_\tau} I_m \succ PC$, and so $PC + \frac{1}{T_\tau} I_m$ will be dominated by $\frac{1}{T_\tau} I_m$. Thus

$$\tilde{C} \rightarrow \frac{1}{T_\tau} \left(\frac{1}{T_\tau} I_m \right)^{-1} C = C$$

as $P \rightarrow 0$. This implies that the capacity would tend to zero too. Therefore, for small P , the number of training symbols T_τ must be large in order for $\frac{1}{T_\tau} I_m$ to be of the same order as PC and so making \tilde{C} in some sense.

Now consider the other extreme, that is when P is large and recall the capacity formula

$$\mathcal{C}(T_\tau) = (T - T_\tau) (\log_2 \det(PC + I_m) - \log_2 \det(P\tilde{C} + I_m))$$

In this case, we get $\frac{1}{T_\tau} I_m \prec PC$. Inspecting the formula for \tilde{C} again, we see that $PC + \frac{1}{T_\tau} I_m$ will be dominated by PC and thus

$$\tilde{C} \approx \frac{1}{T_\tau} (PC)^{-1} C = \frac{1}{T_\tau P}.$$

Hence,

$$\begin{aligned} \log_2 \det(P\tilde{C} + I_m) &\approx \log_2 \det \left(\frac{1}{T_\tau} I_m + I_m \right) \\ &= m \log_2 \left(\frac{1}{T_\tau} + 1 \right) \end{aligned}$$

Note also that for large P , we have that

$$\log_2 \det(PC + I_m) \approx \log_2 \det(PC).$$

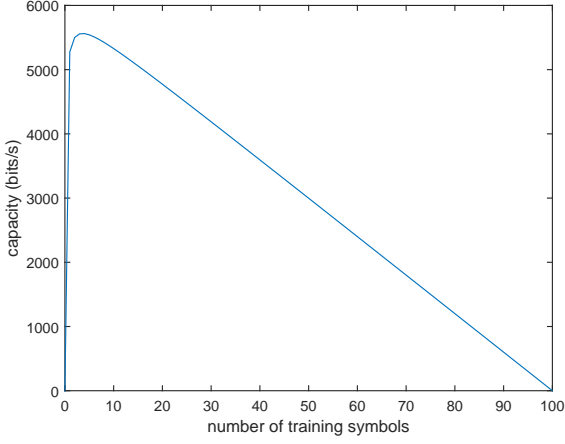


Fig. 3. A plot of the 10×1 channel capacity as a function of the number of training symbols with transmission power given by $P = 100$.

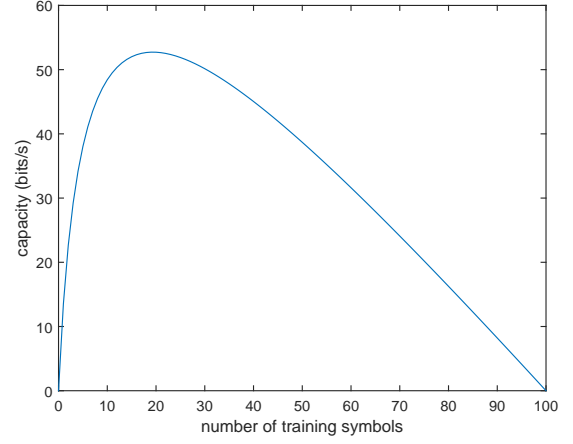


Fig. 4. A plot of the 10×1 channel capacity as a function of the number of training symbols with transmission power given by $P = 0.01$.

So for large P , the capacity may be well approximated by

$$\mathcal{C}(T_\tau) \approx (T - T_\tau)(\log_2 \det(PC) - m \log_2 \left(\frac{1}{T_\tau} + 1 \right))$$

We see that the capacity increases linearly with decreasing T_τ while it increases logarithmically with increasing T_τ . Therefore, the optimal choice of the number of training symbols $T_\tau \rightarrow 1$ as $P \rightarrow \infty$.

VII. CONCLUSION

We considered the problem of deciding the power ratio between the training symbols and data symbols in order to maximize the channel capacity for transmission over uncertain channels with channel state information at the transmitter and receiver. We considered the worst case capacity as a performance measure, where the transceiver maximizes the minimal capacity over all distributions of the measurement noise with a fixed covariance known at both the transmitter and receiver. We presented an exact expression of the channel capacity as a function of the channel covariance matrix, the noise covariance matrix, and the number of training symbols used during a coherence time interval. This expression determines the number of training symbols that need to be used by finding the optimal integer number of training symbols that maximize the channel capacity. We also showed by means of numerical examples the trade-off between the number of training and

data symbols. The results indicate that when the transmission power (or equivalently the signal to noise ratio) is high, a smaller number of training symbols is required to maximize the capacity compared to the low transmission power case. We confirm these observations theoretically considering the asymptotic behavior of the power as it grows large or decreases to very small values. Future work considers the general MIMO case, which is more involved due to combinatorial issues arising in choosing the number of training symbols for different transmitting antennas.

REFERENCES

- [1] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran. Multiuser MIMO achievable rates with downlink training and channel state feedback. *Information Theory, IEEE Transactions on*, 56(6):2845–2866, June 2010.
- [2] G. Fodor, P. Di Marco, and M. Telek. Performance analysis of block and comb type channel estimation for massive mimo systems. In *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*, pages 62–69, Nov 2014.
- [3] B. Hassibi and B. M. Hochwald. How much training is needed in multiple-antenna wireless links? *Information Theory, IEEE Transactions on*, 49(4):951–963, April 2003.
- [4] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.
- [5] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1999.
- [6] T. L. Marzetta and B.M. Hochwald. Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *Information Theory, IEEE Transactions on*, 45(1):139–157, Jan 1999.

APPENDIX

Proof of Theorem 1

Define

$$\mathcal{C}_\star = \sup_{y_d \sim \mathcal{CN}(0, X_d \otimes C + W_d)} \inf_{\substack{v_d \sim \mathcal{CN}(0, V_d) \\ V_d = I_{mT_d} + X_d \otimes \tilde{C}}} \mathcal{I}(x_d; y_d) \quad (23)$$

and

$$\bar{\mathcal{C}} = \sup_{\mathbb{E}\{x_d(t)x_d^*(t)\} = X_d} \inf_{\substack{v_d \sim \mathcal{CN}(0, V_d) \\ V_d = I_{mT_d} + X_d \otimes \tilde{C}}} \mathcal{I}(x_d; y_d), \quad (24)$$

Let the estimator at the receiver be a linear function of the received signal y_d , that is the estimate of x_d is given by $\hat{x}_d = L_d(\hat{H}_d x_d + v_d)$ for some matrix $L_d \in \mathbb{R}^{T_d \times mT_d}$. Since the receiver could be chosen nonlinear, we will get a lower bound on the capacity $\mathcal{C}(T_\tau)$. Introduce $\tilde{x}_d = x_d - \hat{x}_d$. Then

$$\mathcal{C}_\star = \mathcal{I}(x_d; y_d) \quad (25)$$

$$= h(x_d) - h(x_d|y_d) \quad (26)$$

$$= h(x_d) - h(\hat{x}_d + \tilde{x}_d|y_d) \quad (27)$$

$$\geq h(x_d) - h(\tilde{x}_d) \quad (28)$$

Thus,

$$\inf_{\substack{v_d \\ V_d = I_{mT_d} + X_d \otimes \tilde{C}}} \mathcal{I}(x_d; y_d) = h(x_d) - h(\tilde{x}_d) \quad (29)$$

which is attained for v_d such that $y_d = \hat{H}_d x_d + v_d$ is Gaussian, which implies in turn that $\hat{x}_d = L_d(\hat{H}_d x_d + v_d)$ and \tilde{x}_d are Gaussian, and \tilde{x}_d is independent of y_d , so equality holds in (28). Also, $h(x_d)$ is maximized for x_d when it's Gaussian under a fixed covariance. Hence, $\mathcal{C}(T_\tau) \geq \mathcal{C}_\star$. Now consider an arbitrary receiver, that is not necessarily linear and suppose that v_d is Gaussian and independent of x_d . This gives the capacity upperbound $\bar{\mathcal{C}} \geq \mathcal{C}(T_\tau)$. We have that

$$\mathcal{I}(x_d; y_d) = h(y_d) - h(y_d|x_d) \quad (30)$$

$$= h(y_d) - h(v_d|x_d) \quad (31)$$

$$= h(y_d) - h(v_d) \quad (32)$$

where the inequality (32) holds since v_d is assumed to be independent of x_d . Since $h(y_d)$ is maximized when y_d is Gaussian, we get $\mathcal{C}(T_\tau) \leq \bar{\mathcal{C}} = \mathcal{C}_\star$. Since we already have the inequality $\mathcal{C}(T_\tau) \geq \mathcal{C}_\star$, we conclude that $\mathcal{C}(T_\tau) = \mathcal{C}_\star$, and clearly x_d and y_d Gaussian give the worst case capacity $\mathcal{C}(T_\tau) = \mathcal{C}_\star$.

Now let the eigenvalue decompositions of X_d , C , and \tilde{C} be given by $X_d = U\Sigma U^*$, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_{T_d})$, $C = \bar{U}\bar{\Sigma}\bar{U}^*$, $\bar{\Sigma} = \text{diag}(\bar{\sigma}_1, \dots, \bar{\sigma}_{T_d})$, and $\tilde{C} = \tilde{U}\tilde{\Sigma}\tilde{U}^*$,

$\tilde{\Sigma} = \text{diag}(\tilde{\sigma}_1, \dots, \tilde{\sigma}_{T_d})$. Then, the mutual information between the Gaussian input x_d and Gaussian output y_d satisfies

$$\mathcal{I}(x_d; y_d) = h(y_d) - h(y_d|x_d) \quad (33)$$

$$= \log_2 \det(X_d \otimes C + I_m \otimes I_{T_d}) - \log_2 \det(X_d \otimes \tilde{C} + I_m \otimes I_{T_d}) \quad (34)$$

$$= \log_2 \det(X_d \otimes C + I_{mT_d}) - \log_2 \det(X_d \otimes \tilde{C} + I_{mT_d}) \quad (35)$$

$$= \log_2 \det((U \otimes \bar{U})(\Sigma \otimes \bar{\Sigma})(U \otimes \bar{U})^* + I_{mT_d}) - \log_2 \det((U \otimes \tilde{U})(\Sigma \otimes \tilde{\Sigma})(U \otimes \tilde{U})^* + I_{mT_d}) \quad (36)$$

$$= \log_2 \det((U \otimes \bar{U})^*(U \otimes \bar{U})(\Sigma \otimes \bar{\Sigma}) + I_{mT_d}) - \log_2 \det((U \otimes \tilde{U})^*(U \otimes \tilde{U})(\Sigma \otimes \tilde{\Sigma}) + I_{mT_d}) \quad (37)$$

$$= \log_2 \det(\Sigma \otimes \bar{\Sigma} + I_{mT_d}) - \log_2 \det(\Sigma \otimes \tilde{\Sigma} + I_{mT_d}) \quad (38)$$

$$= \log_2 \left(\prod_{i=1}^{T_d} \det(\sigma_i \bar{\Sigma} + I_m) \right) - \log_2 \left(\prod_{i=1}^{T_d} \det(\sigma_i \tilde{\Sigma} + I_m) \right) \quad (39)$$

$$= T_d \log_2 \left(\prod_{i=1}^{T_d} \frac{\det(\sigma_i \bar{\Sigma} + I_m)}{\det(\sigma_i \tilde{\Sigma} + I_m)} \right)^{\frac{1}{T_d}} \quad (40)$$

$$\leq T_d \log_2 \left(\frac{1}{T_d} \sum_{i=1}^{T_d} \frac{\det(\sigma_i \bar{\Sigma} + I_m)}{\det(\sigma_i \tilde{\Sigma} + I_m)} \right) \quad (41)$$

where (36) follows from Proposition 1, (37) follows from Proposition 2, (38) follows from Proposition 1, and (41) follows from Proposition 3, with equality if and only if $\sigma_1 = \sigma_2 = \dots = \sigma_{T_d}$. Since $\sigma_1 + \sigma_2 + \dots + \sigma_{T_d} = \text{Tr}(\Sigma) = \text{Tr}(X) = nP$, we must have

$$\sigma_1 = \sigma_2 = \dots = \sigma_{T_d} = P$$

This implies that the capacity maximizing input covariance is $X = P \cdot I_{T_d}$. Thus, the maximum capacity is given by

$$\mathcal{C}(T_\tau) = \log_2 \det((PI_{T_d}) \otimes (\hat{C} + \tilde{C}) + I_m \otimes I_{T_d}) - \log_2 \det((PI_{T_d}) \otimes \tilde{C} + I_m \otimes I_{T_d}) \quad (42)$$

$$= T_d (\log_2 \det(PC + I_m) - \log_2 \det(P\tilde{C} + I_m)) \quad (43)$$

$$= (T - T_\tau) (\log_2 \det(PC + I_m) - \log_2 \det(P\tilde{C} + I_m)) \quad (44)$$

and the proof is complete.

Supplement for Example 2

The random matrix C used in Example 2 is given by

$$C = \begin{pmatrix} 12.618 & -2.5315 & -2.2424 & 1.1965 & -1.5896 \\ -2.5315 & 8.7639 & 2.0577 & -4.3889 & 2.0117 \\ -2.2424 & 2.0577 & 6.0997 & 0.2384 & -1.1894 \\ 1.1965 & -4.3889 & 0.2384 & 7.327 & 1.1523 \\ -1.5896 & 2.0117 & -1.1894 & 1.1523 & 10.2643 \\ 2.5086 & -0.011 & -1.6082 & 0.3583 & 1.3222 \\ -4.5906 & -0.499 & -4.5764 & 0.705 & -0.1717 \\ -1.398 & 3.1713 & -2.319 & -5.3612 & -2.361 \\ 1.9345 & 2.1956 & -0.6284 & -2.2747 & -0.6889 \\ -4.0798 & 0.2636 & -0.5846 & -0.7751 & 1.2117 \\ 2.5086 & -4.5906 & -1.398 & 1.9345 & -4.0798 \\ -0.011 & -0.499 & 3.1713 & 2.1956 & 0.2636 \\ -1.6082 & -4.5764 & -2.319 & -0.6284 & -0.5846 \\ 0.3583 & 0.705 & -5.3612 & -2.2747 & -0.7751 \\ 1.3222 & -0.1717 & -2.361 & -0.6889 & 1.2117 \\ 2.1366 & 0.2868 & -0.8628 & 1.2528 & -1.1311 \\ 0.2868 & 18.4323 & 10.9609 & 2.3883 & 2.0394 \\ -0.8628 & 10.9609 & 20.4969 & 10.5705 & 1.3217 \\ 1.2528 & 2.3883 & 10.5705 & 9.7425 & -0.307 \\ -1.1311 & 2.0394 & 1.3217 & -0.307 & 3.3511 \end{pmatrix}$$